VISUAL DISTINCTNESS DETERMINED BY PARTIALLY INVARIANT FEATURES.

J.A. Garcia, J. Fdez-Valdivia
Departamento de Ciencias de la Computacion e I.A.
Univ. de Granada. E.T.S. de Ingenieria Informatica.
18071 Granada. Spain
E-mail: jags@decsai.ugr.es, J.Fdez-Valdivia@decsai.ugr.es

Xose R. Fdez-Vidal
Departamento de Fisica Aplicada.
Univ. de Santiago de Compostela. Facultad de Fisica.
15706 Santiago de Compostela. Spain
E-mail: faxose@usc.es

Rosa Rodriguez-Sanchez
Departamento de Informatica. Universidad de Jaen
Escuela Politecnica Superior. 23071 Jaen. Spain
E-mail: rosa@ujaen.es

1. SUMMARY

This paper describes a system for the automatically learned partitioning of "visual patterns" in digital images, based on a sophisticated, band-pass, filtering operation, with fixed scale and orientation sensitivity. The "visual patterns" are defined as the features which have the highest degree of alignment in the statistical structure across different frequency bands. Here we show a computational visual distinctness measure computed from the image representational model based on visual patterns. It is applied to quantify the visual distinctness of targets in complex natural scenes. We also investigate the relation between the computational distinctness measure and the visual target distinctness measured by human observers.

2. INTRODUCTION

Images issued from the environment should not be presumed to be random patterns. Instead, real-world images contain characteristic statistical regularities that set them apart from purely random images. There are a number of statistical properties that we might consider when looking at real-world images, and many of the important forms of structure that are contained in 2D images require higher-order statistics characterization. Moreover, Field [1] noted that there is likely to be a variety of features which extend across different frequency bands. For instance, the presence of edges and lines in an image corresponds to a type of congruence between the different scales of the image which is destroyed when the phases are randomized [2]. These features exist because some degree of alignment exists between the phases at different frequencies. There are also other forms of congruence across scales in 2D digital images. Field [3] suggested that the power spectra of natural images falls off as a function of frequency by a factor of approximately $1/k^2$. This implies that the image will have constant variance across scales: the contrast as measured by the variance in pixel intensities should remain roughly constant, independently of the viewing distance.The perceptual organization capabilities of human vision seem to exhibit the properties of detecting viewpoint-invariant structures and calculating varying degrees of significance for individual instances [4]. Lowe [5] proposed that the structures to be detected in the image should be formed bottom-up using perceptual grouping operations that exhibit exactly these properties in the absence of domain knowledge, yet must be of sufficient specificity to serve as indexing terms into a database of objects. Given that we often have no priori knowledge of viewpoint for the objects in a database, these indexing features that are detected in the image must reflect properties of the objects that are at least partially invariant over a wide range of viewpoints of some corresponding three-dimensional structure. This means that it is useless to look for features with particular sizes or orientations or other properties that are highly dependent upon viewpoint. The second constraint on these indexing features is that there must be some way to distinguish the relevant features from the dense background of other image features which could potentially give rise to false instances of the structures.

Often implicit in the interpretation of visual search tasks is the assumption that the detection of targets is determined by the feature-coding properties of low-level visual processing [6]. Instead of assuming that perceived shapes are simple or statistical structure at a particular scale, we think it more appropriate to regard them as "visual patterns", distinguished at an object level.

Here we show a particular scheme for filtering observed images, designed to the automatically learned partitioning of features (visual patterns) which have the highest degree of alignment in statistical structure across different frequency bands. These features are likely to be invariant over a range of scales and orientations and can be judged unlikely to be accidental in origin even in the absence of specific information regarding which objects may be present. Then, we present a computational visual distinctness measure computed from the image representational model based on visual patterns. This measure applies a simple decision rule to the distances between segregated visual patterns, and it will be used to quantify the visual distinctness of targets in complex natural scenes. The analysis to the automatically learned partitioning of "visual patterns" (it has been termed RGFF model) follows three stages: Preattentive stage, Integration stage, and Learning stage. Fig. 1 shows a general diagram describing how the data flows through the RGFF model. This diagram illustrates the analysis on a given image of a military vehicle in a complex rural background.

In the preattentive stage of the RGFF system (Section 3), the clumps of energy in the Fourier spectrum of the image are captured into a collection of oriented spatial-frequency channels, as illustrated in Fig. 1. The segregation of these clumps of energy induces the selection of a subset of activated filters (which are selectively sensitive to them) from a filter bank of log-Gabor functions centered at 12 orientations and 5 ranges. Due to conjugate symmetry, the filter design is only carried out on half the 2D frequency plane. The activated log-Gabor filters produced by the preattentive stage are illustrated in the diagram by ellipses drawn, in the 2D spatial-

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

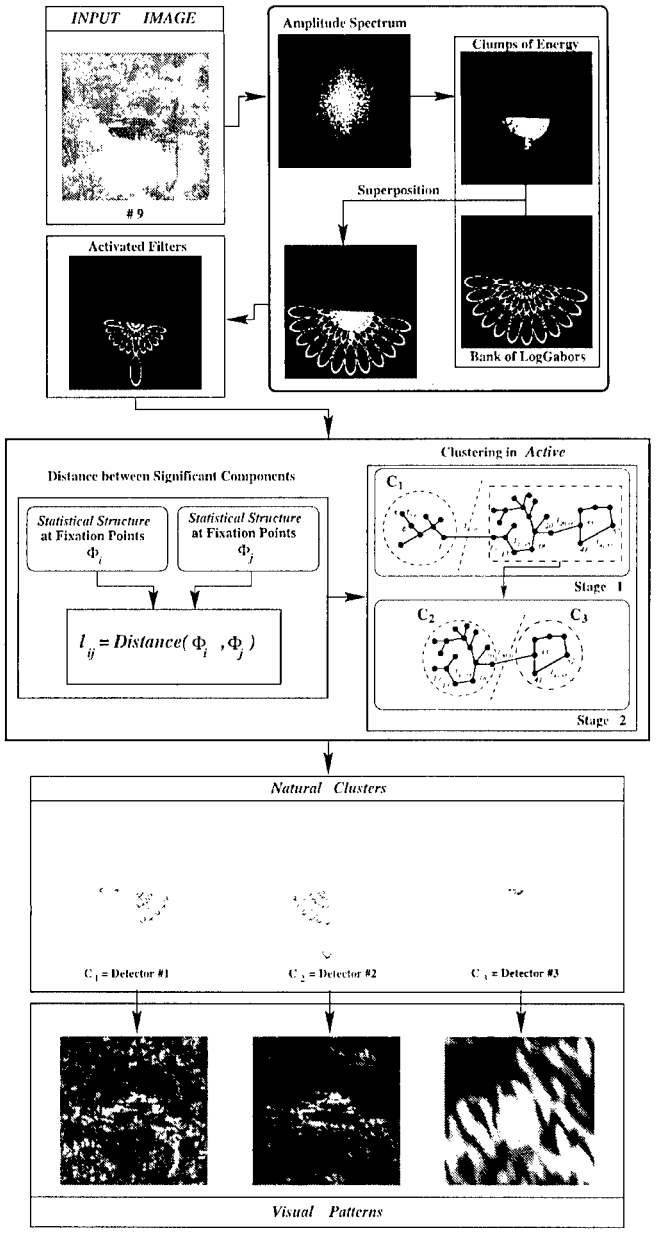


Figure 1: A general diagram describing how the data flows through the representational model..

frequency plane, at the point where their amplitude has decreased to the ($e^{-1/2}$) half width its maximum.
In the integration stage (Section 4), for any two activated filters, their responses are compared based on the distance (a β-norm between their statistical structure, computed over those pixels which form "fixation points" of the filters (local energy peaks on the filtered response).
In the learning stage (Section 5), clustering on the basis of the distance between the activated filters is performed to highlight scale and orientation invariance of responses.
As shown in Fig. 1, three collections of filters were obtained in the Learning stage for the input image in accordance with a constraint of invariance in statistical structure across frequency bands. The filtered responses of activated log-Gabors in each one of the three groupings were summed for the automatic learned partitioning of the visual patterns.
The performance of this notion of visual pattern to segregate potential targets can be visually evaluated in Fig. 1, at the bottom. The dominant signal in the output from detector #2 is the military vehicle (target) which is well preserved. On the contrary, both large structures and fine detail of the natural background were removed, even though significant background clutter that can affect the target distinctness is still present. In fact, the fine details of the natural background, which are not significant for quantifying the target distinctness, are isolated in the output from detector #1. And the lower frequency texture of the background is segregated into the output from detector #3. Fig. 2 demonstrates the ability of the same model to achieve signal separation from superposition of objects on three synthetic images. The image in Fig. 2.A1 was partitioned into two "visual patterns", as shown in Figs. 2.C1 and 2.E1. In the learning stage, the set of activated filters was partitioned into two groupings of filters, as shown in Figs. 2.B1 and 2.D1. The "visual pattern" shown in Fig. 2.C1 (respectively, Fig. 2.E1) was obtained by the sum of the responses over filters in Fig. 2.B1 (resp., Fig. 2.D1). The "visual patterns" obtained by the model on Fig. 2.A2, are illustrated in Figs. 2.C2 and 2.E2. The learning stage produced two collections of filters as shown in Figs. 2.B2 and 2.D2. The right column in Fig. 2 shows the signal separation achieved by the analysis on the input image given in Fig. 2.A3.
Finally, Section 6 presents the computational visual distinctness measure computed from the image representational model based on visual patterns. As illustrated in Fig. 10, this measure applies a simple decision rule to the distances between segregated visual patterns.

3. PREATTENTIVE STAGE

In the RGFF model, the encoding strategy will rely on the combined activity of subsets of filters. Only a small number of units will contribute to the detection of each visual pattern. These collections of filters will be derived from a learning stage, based on the degree of congruence between the responses of strongly responding filters that a preattentive stage produces. There are two basic assumptions for this first stage:

1. Spatial information on the image is analyzed by multiple filters, each of which is sensitive to patterns whose spatial frequencies are in a particular range.
2. The RGFF model bases its responses only on those filters sensitive to relevant forms in the complex scene.

These assumptions are in agreement with models of spatial-frequency channels which are quite successful for the detection of visual patterns [7]. The output of the preattentive stage will be the units from a fixed filter bank of log-Gabors which are tuned to the clumps of energy in the Fourier spectrum of the given image. The selected units are the filters in the bank which strongly respond to some pattern that the image contains. These filters are referred as the "activated" filters of the bank. Also for each activated filter, pixels whereupon the focus of attention should be shifted to measure congruence and which form "fixation points", are computed as local energy peaks on the filtered response. This processing is based on current models of human visual search and detection which assume that a preattentive stage indicates potentially interesting image regions, and where a serial stage is deployed to analyze them in detail [6,7].

3.1. Bank of filters

The set of filters used in the decomposition of the picture consists of log-Gabor filters of different spatial frequencies and orientations [3]. Log-Gabor functions, by definition, have no DC component. The transfer function of the log-Gabor has extended tails at the high frequency end. Thus it should be able to encode natural images more efficiently than ordinary Gabor functions, which would over-represent the low-

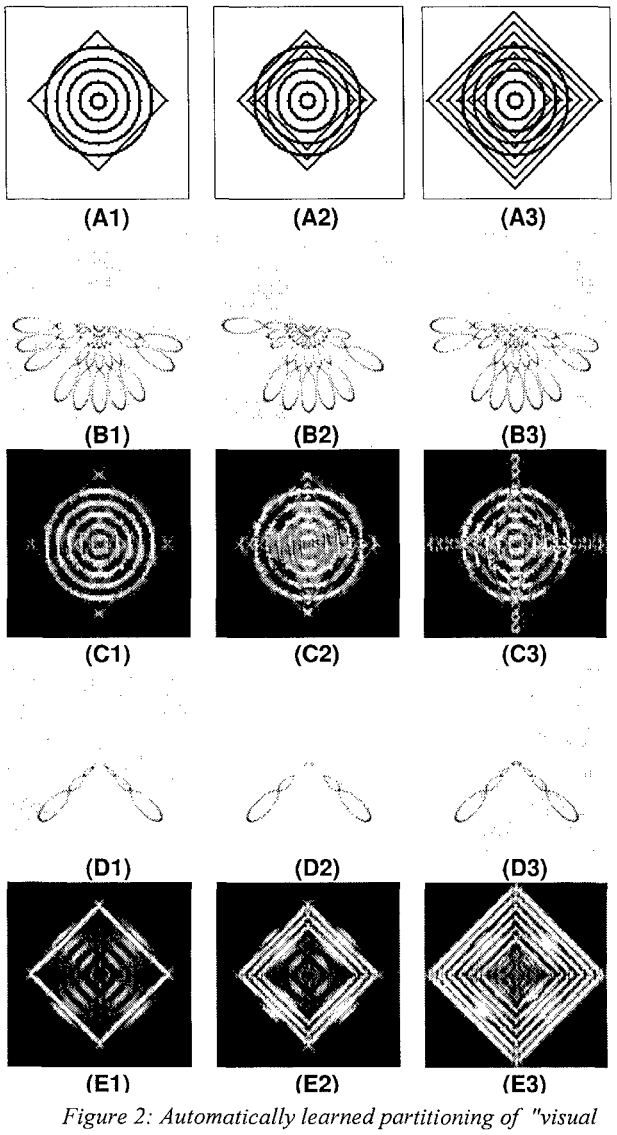


Figure 2: Automatically learned partitioning of "visual patterns" in synthetic image data.

frequency components and under-represent the high frequency components in any encoding process. Another argument in support of the log-Gabor functions is the consistency with measurements on the mammalian visual system [8].

A Log-Gabor filter determines a Gaussian in the spatial frequency domain around some central frequency (r_o, θ_o). It can be represented in the frequency domain as the sum of the even-symmetric log-Gabor filter and i times the odd-symmetric log-Gabor filter as follows:

$$\phi(r_o,\theta_o) = \exp\left\{-\frac{(\log(r/r_o))^2}{2(\log(\sigma_r/r_o))^2}\right\}\exp\left\{-\frac{(\theta-\theta_o)^2}{2\sigma_\theta^2}\right\} \qquad (1)$$

where θ_o is the orientation angle of the filter, r_o is the central radial frequency, σ_θ and σ_r are the angular and radial sigma of the Gaussian, respectively.

The convolution of a log-Gabor function (whose real and imaginary parts are in quadrature) with a real image results in a complex image. Its norm is called energy and its argument is called phase. The local energy of the image analyzed by a log-Gabor filter (hereafter, filtered response) can be expressed as [3]:

$$E(x,y) = \sqrt{O_{even}^2(x,y) + O_{odd}^2(x,y)} \qquad (2)$$

where $O_{even}(x,y)$ is the image convolved with the even-symmetric log-Gabor filter and $O_{odd}(x,y)$ is the image convolved with the odd-symmetric log-Gabor filter.

The real-valued function given in equation (1) can be multiplied by the frequency representation of the image and after transforming the result back to the spatial domain, the results of applying the oriented energy filter pair are extracted as simply the real component for the even-symmetric filter and the imaginary component for the odd-symmetric filter [9]. The bank of the filters should be designed so that it tiles the frequency plane uniformly (the transfer function should be a perfect bandpass function). The length to width ratio of the filters controls their directional selectivity. The ratio can be varied in conjunction with the number of orientations used in order to achieve a coverage of a 2D spectrum. Furthermore, as the degree of blurring introduced by the filters increases with their orientational selectivity, they must be carefully chosen to minimize the blurring. Hence we consider a filter bank with the following features:

1. The spatial frequency plane is divided into 12 different orientations.
2. The radial axis is divided into 5 equal octave bands. In a band of width 1 octave, spatial frequency increases with a factor 2. The highest filter (for each direction) is positioned near the Nyquist frequency to avoid ringing and noise. The wavelength of the five filters in each direction is set at 3, 6, 12, 24, and 48 pixels, respectively.
3. The radial bandwidth is chosen as 1.2 octaves.
4. The angular bandwidth is chosen as 15 degrees.

Twelve different angles for each resolution are chosen and five different resolutions are used. The resultant filter bank is illustrated in Fig.1. Due to conjugate symmetry, the filter design is only carried out on half the 2D frequency plane. The log-Gabor filters are illustrated in the diagram by ellipses drawn, in the 2D spatial-frequency plane, at the point where their amplitude has decreased to the $(e^{-1/2})$ half width its maximum.

3.2. Activated filters in the bank

In order to decompose the image into its most significant components, strongly responding filters should be selected for the input image.

Let *Active* be the set of filters in the bank that strongly respond to the spatial information content. They will be selectively sensitive to patterns in the scene. These patterns produce clumps of energy upon the Fourier spectrum of the image. The activated units from the bank are then simply those filters whose amplitude spectrum and some clump of energy in the image amplitude spectrum overlap to some extent, as illustrated in Fig.1.

3.3. Selection of fixation points

In the integration stage, for any given two activated filters, a distance between them is derived via distances between their statistics. The distance chosen is the β-norm, computed over those pixels which form "fixation points" of the filters. The "fixation points" are simply local energy peaks on the filtered response. The standard argument for selecting regions of high Gabor energy is that they would provide a good starting point for exploring common grounds between several activated filters in the Gabor space. The implementation of the local-energy model used here is the one presented in [10]. Given the original image, the local energy map E_i for the activated filter ϕ_i, given in equation (2), yields a representation in the

space spanned by two functions, $O_{even}(x,y)$ and $O_{odd}(x,y)$, where $O_{even}(x,y)$ is the image convolved with the even-symmetric log-Gabor filter and $O_{odd}(x,y)$ is the image convolved with the odd-symmetric log-Gabor filter at (x,y). Hence, the detection of peaks on the E_i map acts as a detector of significant features on the filtered response.

4. INTEGRATION STAGE

Given a decomposition of the original image into its most significant components, only a further element is needed to define the concept of visual pattern: a distance measure, denoted as $Distance(\phi_i, \phi_j)$, between the statistical structures of the filtered responses for each pair of filters ϕ_i and ϕ_j (Section 4.2.). Then, $Distance(\phi_i, \phi_j)$ returns a value of the degree of congruence between statistical structure at different scales and orientations.

There are two basic assumptions for measuring congruence between two filtered responses in this second stage:

1. The similarity between two filtered responses can be measured by the Quick pooling of the differences between their statistical structure.
2. The measure of similarity is not simply computed globally over the entire filtered response, but semi-locally at locations that are local energy peaks (fixation points).

Previously, it was demonstrated [11] that a measure based on these two assumptions produces a good predictor of target saliency for humans performing visual search and detection tasks.

4.1. Definition of integral feature for the partitioning of visual patterns

For each activated filter ϕ_i, the respective filtered response may be represented by any subset of the following separable features:

1. The phase value defined as:

$$T_1^i(x,y) = arctan\frac{O_{even}(x,y)}{O_{odd}(x,y)} \tag{3}$$

where $O_{even}(x,y)$ is the image convolved with the even-symmetric log-Gabor filter of ϕ_i, and $O_{odd}(x,y)$ is the image convolved with the odd-symmetric log-Gabor filter of ϕ_i at (x,y) (Section 3.1.).

2. A normalized measure of local energy as given by:

$$T_2^i(x,y) = \frac{E_i(x,y)}{\sum_{\{j/\phi_j \in Active\}} E_j(x,y)} \tag{4}$$

where $E_i(x,y)$ denotes the local energy at (x,y) for filter ϕ_i (see equation 2 for further details), and *Active* is the set of activated filters for the image. This definition of a normalized local energy incorporates lateral interactions among activated filters to account for between-filter masking.

3. The local standard deviation of the normalized local energy defined as:

$$T_3^i(x,y) = \left(\frac{1}{Card\ [W(x,y)]} \sum_{(p,q)\in W(x,y)} (T_2^i(p,q) - \mu)^2 \right)^{1/2} \tag{5}$$

where

$$\mu = \frac{1}{Card\ [W(x,y)]} \sum_{(p,q)\in W(x,y)} T_2^i(p,q)$$

and $T_2^i(p,q)$ as given in equation (4). The neighborhood $W(x,y)$ is defined as the set of pixels contained in a disk of radius r centered at (x,y). Let r be defined as the Euclidean distance between (x,y) and the nearest local minimum to (x,y) on the energy map E_i. Since the nearest local minimum to (x, y) on the local energy map marks the beginning of another potential structure, our selection for the neighborhood $W(x, y)$ avoids interference with such a structure while the local variation is computed [10].

4. The local contrast of the normalized local energy defined as:

$$T_4^i(x,y) = \frac{T_3^i(x,y)^2}{\mu} \tag{6}$$

where

$$\mu = \frac{1}{Card\ [W(x,y)]} \sum_{(p,q)\in W(x,y)} T_2^i(p,q)$$

5. The local entropy of the normalized local energy within $W(x,y)$, noted as $T_5^i(p,q)$.

Although we propose these five features, any other intent to capture relevant characteristics of the scene, while stable for the representation of the image is also conceivable. Hereafter, *an "integral feature" is defined as a particular subset of separable features at a fixation point* [12].

For representing the filtered responses of the input image, different definitions of integral feature can be given based on different subsets of separable features. Consequently, the system should learn the best integral feature definition for the input image in which to look for invariance across orientations and scales. This point is analyzed in Section 6.4.

4.2. Congruence in integral features between two filtered responses

In order to define a distance between the integral features of two filtered responses, we need to specify how the differences in each separable feature are to be pooled into an overall difference at fixation points.

Let ϕ_i and ϕ_j be a pair of activated filters in *Active*. Let $T^i(x,y)=(T_{l_k}^i(p,q))_{1\le k\le L}$, with $l_k \in \{1,2,...,5\}$, be the integral feature at (x,y) computed on the filtered response of ϕ_i, based on a number of L separable features (Section 4.1.). In a similar way, let $T^j(x,y) = (T_{l_k}^j(p,q))_{1\le k\le L}$, be the integral feature at (x,y) on the filtered response of ϕ_j.

We take $D[T^i(x,y), T^j(x,y)]$ defining a distance measure between integral features $T^i(x,y)$ and $T^j(x,y)$ as given by the equation:

$$D[T^i(x,y), T^j(x,y)] = \sum_{k=1}^{L} \frac{1}{Max_{l_k}} d(T_{l_k}^i(x,y), T_{l_k}^j(x,y)) \tag{7}$$

where normalization Max_{l_k} is defined as:

$$Max_{l_k} = \max_{n\ m\ \phi_n,\phi_m \in Active} \left\{ d(T_{l_k}^n(p,q), T_{l_k}^m(p,q)) \mid (p,q) \in FP(n) \right\}$$

with $FP(n)$ being the fixation points for the activated filter ϕ_n and *Active* being the set of activated filters; and where for $l_k=1$, we have:

$$d(T_1^i(x,y), T_1^j(x,y)) = \left| arctan\frac{sin(T_1^i(x,y) - T_1^j(x,y))}{\cos(T_1^i(x,y) - T_1^j(x,y))} \right| \tag{8}$$

and for $l_k = 2,3,4,5$:

$$d(T_{l_k}^i(x,y), T_{l_k}^j(x,y)) = \left| T_{l_k}^i(x,y) - T_{l_k}^j(x,y) \right| \tag{9}$$

The congruence in integral features between two filtered responses is computed by using Quick pooling [13]. It is the most common model of integration over spatial extent, and is

essentially the square root of the squares sum except that the exponent is not restricted to the value of 2. The Quick pooling can be viewed as a metric in a multidimensional space, and it is sometimes known as Minkowski metric.

The distance between the filtered responses of ϕ_i and ϕ_j, which provides a measure of the extent to which features extend through frequency, is given by:

$$Distance(\phi_i,\phi_j)=Dist^2[i,j]+Dist^2[j,i] \tag{10}$$

where:

$$Dist[p,q]=\frac{1}{Card[FP(p)]}\left(\sum_{(x,y)\in FP(p)}\left|D[T^p(x,y),T^q(x,y)]\right|^\beta\right)^{\frac{1}{\beta}} \tag{11}$$

with $FP(p)$ being the set of fixation points for the activated filter ϕ_p; and where $D[T^p(x,y), T^q(x,y)]$ is defined as given in equation (7).

The default value of the exponent β in equation (11) is 3. Graham [7] discussed at some length several interpretations of the Quick pooling formula and the selection of the pooling exponent.

5. LEARNING STAGE

Based on a measure of the extent to which features extend through frequency, noted as Distance(ϕ_i,ϕ_j), a "visual pattern" is simply defined as congruence in statistical structure, as measured by Distance, across a range of 2D spatial frequency bands.

The individual filters spanning this particular range of bands will determine a natural cluster of units, noted as C_n, in the set of activated logGabors *Active*. By taking into account the statistical congruence across this range of frequency bands, a pair of filters ϕ_i and ϕ_j will belong to the same natural cluster C_n if there exists certainly continuity (i.e., there exists similarity in some statistics at the same spatial locations) across the filtered responses for an intermediate sequence of filters, between ϕ_i and ϕ_j, in C_n.

Therefore the definition of "visual pattern" induces a partition in *Active* into a number of natural clusters $C_1, C_2, ..., C_N$ such that:

$$Active=\bigcup_{n=1}^{N}C_n \text{ , and } C_p\cap C_q=\phi, \quad with \quad p\neq q, p,q=1,2,..,N \tag{12}$$

where, for each C_n, a pair of filters ϕ_i , $\phi_j \in C_n$ if there exists a sequence of filters $\phi_{n_1},\phi_{n_2},...,\phi_{n_l}$ in C_n such that

$$\begin{aligned} &Distance\ (\phi_i,\phi_{n_1})\le\varepsilon_n \\ &Distance\ (\phi_{n_l},\phi_j)\le\varepsilon_n \\ &Distance\ (\phi_{n_k},\phi_{n_{k+1}})\le\varepsilon_n, k=1,2,...,l-1 \end{aligned} \tag{13}$$

where ε_n denotes the degree of statistical congruence between a pair of filters in C_n and verifies that:

$$\begin{aligned} &Distance\ (\phi_p,\phi_q)>\varepsilon_n, \\ &\forall\,\phi_p,\phi_q:\phi_p\in C_n,\phi_q\in Active-C_n \end{aligned} \tag{14}$$

The clustering of activated filters is performed as described in Section 5.1. Figs. 3 and 4 illustrate the performance of the clustering on several images of a target in a complex rural background.

The image in Fig. 3.A1 was partitioned into the two visual patterns shown in Figs. 3.C1 and 3.E1.

In the clustering process, the set of activated filters was partitioned into two collections of filters, as shown in Figs. 3.B1 and 3.D1. The "visual pattern" shown in Fig. 3.C1 (respectively, Fig. 3.E1) was obtained by the sum of the responses over filters in Fig. 3.B1 (resp., Fig. 3.D1).

The right column in Fig. 3 shows the separation achieved by the analysis on the image in Fig. 3.A2.

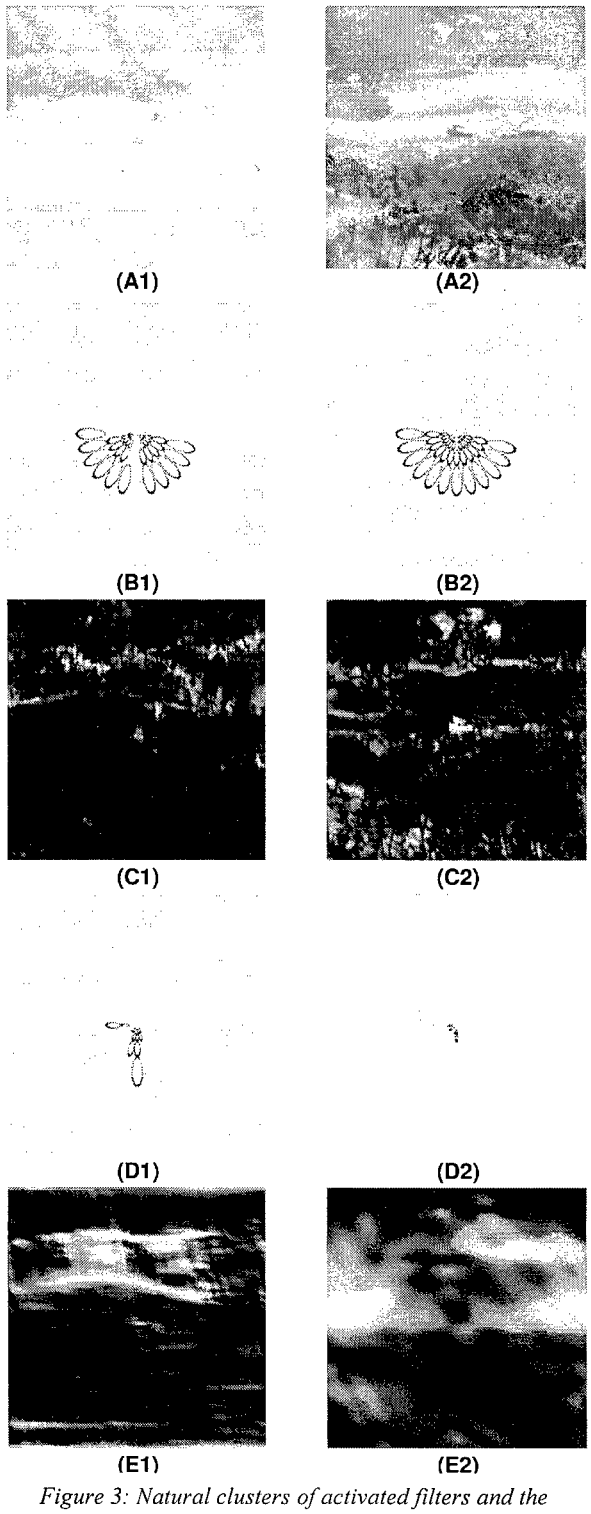


Figure 3: Natural clusters of activated filters and the respective visual patterns.

The visual patterns produced by the model on Fig. 4.A1 (respectively, 4.A2) are illustrated in Figs. 4.C1, 4.E1, and 4.G1 (resp., Figs. 4.C2, 4.E2, and 4.G2). In both cases, the clustering of activated filters produced three collections of filters as shown in Fig. 4.

5.1. Clustering of activated filters

We formulate the problem as the clustering of a dataset $X=\{i \mid \phi_i \in Active\}$ into a number N of natural clusters $\zeta_0, \zeta_1, ..., \zeta_{N-1}$. We call clusters natural if the membership is determined fairly well in a natural way by the data.

This clustering is reduced to a sequence of stages of simpler partitioning [14]. At each stage j, a subset X^j of X is divided into only two classes (for $j=0$, $X^0 = X$):
1. a natural cluster ζ_j which contains all the data points (filters) in X^j which are assigned the same class of a seed point (filter) $seed_j$, with $seed_j$ being picked randomly from X^j, and
2. the data, X^j-ζ_j, still not placed in any existing cluster, noted as $\zeta_0, \zeta_1, ..., \zeta_{j-1}$.

The clarity of separation between clusters, as measured by a dissimilarity function, will be the criterion by which we derive a natural cluster ζ_j at stage j. The dissimilarity function is defined in Section 5.1.1. The criterion by which we define a natural cluster at stage j is presented in Section 5.1.2.
The dynamic process of clustering is stopped at stage j if the class X^j-ζ_j is the empty set. Otherwise, the process progresses, and the subset X^{j+1} to be partitioned at the stage $j+1$, it will be the one defined as $X^{j+1}=X^j$-ζ_j. Finally, the natural clusters in *Active* verifying equations (12)-(14) are induced as:

$$C_n = \{\phi_i \in Active \mid i \in \zeta_{n-1}\} \quad with \quad n = 1.2..., N \qquad (15)$$

and where N denotes the number of clusters into which $X=\{i \mid \phi_i \in Active\}$ was partitioned, that is $\zeta_0, \zeta_1, ..., \zeta_{N-1}$. See Fig. 1 for further illustration of this analysis.

5.1.1. Dissimilarity function

Let X^j be a subset of data not absorbed in any of the existing clusters $\zeta_0, \zeta_1, ..., \zeta_{j-1}$, at the stage j of the dynamic processing; with X^0 being the given data set.
$X^0 = X=\{i \mid \phi_i \in Active\}$. Next we define a graph $GRAPH^j = (X^j, U^j)$ corresponding to the data subset X^j, and with U^j being the set of arcs $u=(k,l)$ between pairs of points in X^j. We associate with each arc $u \in U^j$ a real number $l(u) \geq 0$, and if $u=(k,l)$, we shall also use the notation l_{kl} for $l(u)$. Let l_{kl} be the distance from k to l defined as:

$$l_{kl} = Distance \quad (\phi_k, \phi_l) \qquad (16)$$

where $Distance(\phi_k, \phi_l)$ measures the distance between the filtered response of filters ϕ_k and ϕ_l as given in equation (10). The cost of a path is defined as the greatest distance between two successive vertices on the path. Let $\mu(seed_j,k)$ be a set of arcs constituting a path between two points $seed_j$ and k in X^j. And let $l(\mu)$ represent the cost of $\mu(seed_j,k)$ from $seed_j$ to k defined as follows:

$$l(\mu) = max\{l(u) \mid u \in \mu(seed_j,k)\} \qquad (17)$$

Taking into account that two filters belong to the same cluster if there exists continuity (i.e., there exists similarity in their statistics at the same spatial locations) across the responses of filters in a path between them, the dissimilarity function is next defined as the cost of the optimum path from a seed point to each other on the graph. The optimum path between two data points $seed_j$ and k is the path $\mu^*(seed_j,k)$ from $seed_j$ to k whose maximum *cost* $l(\mu^*)$ is minimum:

$$\mu^*(seed_j,k) = \underset{\mu}{Argmin} \left[max\{l(u) \mid u \in \mu(seed_j,k)\} \right] \qquad (18)$$

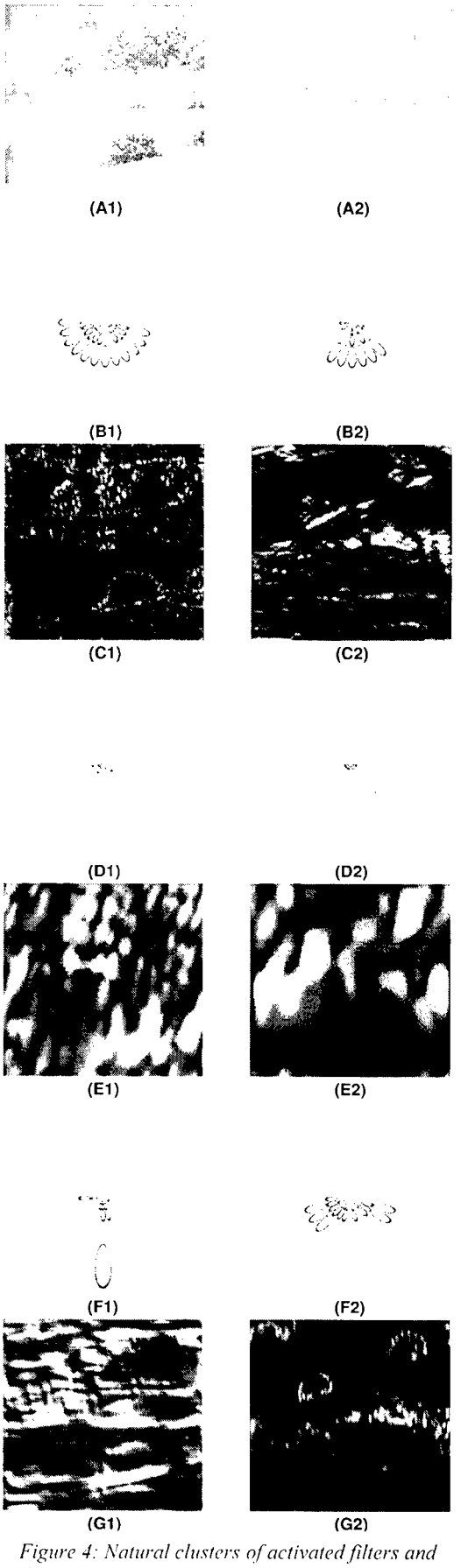


Figure 4: Natural clusters of activated filters and the respective visual patterns.

Visual target distinctness						
Image Pair	R_{P_d}		RMSE		VP_T	
	Value	Rank	Value	Rank	Value	Rank
# 16	96.39	1	2.55	4	4.799	1
# 9	96.27	2	16.90	1	4.531	2
# 37	96.19	3	16.12	2	3.815	3
# 6	90.95	4	1.94	7	3.637	4
# 30	90.66	5	2.37	5	2.713	6
# 26	90.06	6	1.71	8	2.969	5
# 29	89.47	7	1.60	9	2.234	9
# 3	80.02	8	2.83	3	2.254	8
# 21	73.84	9	1.35	10	2.046	10
# 11	73.40	10	1.96	6	2.690	7
P_{cc}	-		0.5		0.8	

Table 1: Comparative results of the RMSE metric and the computational visual distinctness measure.

Hence, the dissimilarity from the viewpoint of $seed_j$ to each k, is defined as the cost of the optimum path $\mu^*(seed_j,k)$ from $seed_j$ to each k:

$$d^{GRAPH^j}(seed_j,k)=l(\mu^*)=max\{l(u)|u\in\mu^*(seed_j,k)\} \quad (19)$$

with μ^* being the optimum path between $seed_j$ and k. The optimal path algorithm is given in [14].

5.1.2. Clarity of separation at stage j

Here we introduce the criterion by which we define the natural cluster ζ_j at stage j.

The set $\{d^{GRAPH^j}(seed_j, k), \text{with } k\in X^j\}$ is firstly ordered to obtain a new function:

$$d_j(i)=d^{GRAPH^j}(seed_j,k_i),\ such\ that\ d_j(i)\le d_j(i+1) \quad (20)$$

where $d_j(i)$ denotes the cost of the optimum path from $seed_j$ to k_i.

Let ε_j represent the degree of closeness that is required between a pair of points that belong to the natural cluster of $seed_j$, noted as ζ_j. Taking into account that $d_j(i)$ measures the closeness between $seed_j$ and k_i with $k_i \in X^j$, we have that ε_j can be defined as:

$$\varepsilon_j = d_j(i^*) \quad (21)$$

with i^* being the location of the first significant rise in the value of $d_j(i)$ when i increases. The value of i^* is computed as the first zero crossing of the second derivative of d_j, as described in Appendix.

A point k_i from X^j is then assigned the same cluster of $seed_j$ if the closeness between $seed_j$ and k_i is less than or equal to ε_j:

$$k_i \in \zeta_j \text{ , if } d_j(i) \le \varepsilon_j;\ otherwise\ \ k_i \notin \zeta_j$$

6. PREDICTING VISUAL TARGET DISTINCTNESS

This section presents a computational visual distinctness measure computed from the image representational model based on visual patterns.

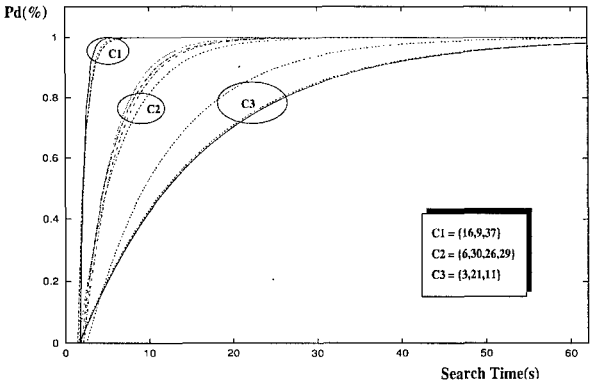


Figure 5: Cumulative distribution functions to the search times for the target scenes

The approach is as follows. First, a psychophysical experiment is performed in which observers estimate the visual distinctness of the target in each of 44 different test scenes (Section 6.2.). Second, a computational measure is defined and then applied to quantify the visual distinctness of the targets (Section 6.3.). Finally, an experiment is performed to investigate the relation between the computational distinctness measure and the visual target distinctness measured by human observers (Section 6.4.).

6.1. Images

The images used in this study are slides made during the DISSTAF (Distributed Interactive Simulation, Search and Target Acquisition Fidelity) field test, that was designed and organized by NVESD (Night Vision & Electro-optic Sensors Directorate, Ft. Belvoir, VA, USA) and that was held in May and June 1995 in Fort Hunter Liggett, California, USA [15].These slides depict 44 different scenes.

Each scene represents a military vehicle in a complex rural background.The 9 different vehicles that are deployed as search targets are respectively a BMP-1, a BTR-70, an HMMVV-Scout, a HMMVV-Tow, an M1A1, an M3-Bradley, an M60, an M113, and a T72. The visibility of the targets varies throughout the entire stimulus set. This is mainly due to variations in the structure of the local background, the viewing distance, the luminance distribution over the target support (shadows), the orientation of the targets, and the degree of occlusion of the targets by vegetation.

The images used in the computational experiments are subsampled to 256x256 pixels. For each scene t, containing a target (vehicle), a corresponding empty scene e was created [6].The empty scene is everywhere equal to the target scene, except at the location of the target, where the target support is filled with the local background.This replacement is done by hand, using the rubber stamp tool in Photoshop 3.05.The result is judged by eye and is accepted if the variation in the background over the target support area does not appear to have an appreciable contrast with the natural variation in the local background.

In the experiment here reported (Section 6.4), the digital images were (see Figs. 6-9): (i) ten complex natural images containing a single target that correspond to the scenes 16, 9, 37, 6, 30, 26, 29, 3, 21, and 11, from the 44 slides made during the DISSTAF field test; and (ii) the corresponding empty images of the same rural backgrounds without target, that were created using the rubber stamp tool in Photoshop 3.05.

For each target image, Figs. 6-9 illustrate the simple thresholding of the visual pattern produced by the natural cluster of filters in *Active* that segregates the military vehicle

(target detector). Simple thresholding was applied to remove small response values which were present in the output of the target detector. In the same figures, it is also shown the visual pattern's thresholding produced by the target detector when it is applied on the respective image without target.
To produce the results shown in Figs. 6-9, the definition of integral feature used for the partitioning of the visual patterns in accord with a constraint of invariance was as follows (Section 4.1):

- for the target scenes 30, 37, and 16, $T = (T_1, T_3, T_5)$;
- for 26 and 6, $T = (T_5)$;
- for 9, $T = (T_1, T_2, T_4)$;
- for 29, $T = (T_1, T_4, T_5)$;
- for 3, $T = (T_3)$;
- for 21, $T = (T_4, T_5)$; and
- for 11, $T = (T_1, T_2)$.

Section 6.4 analyzes how the best definition of integral feature for predicting visual target distinctness can be estimated on a dataset of example.

6.2. Psychophysical target distinctness

A psychophysical experiment was performed in which observers estimate the visual distinctness of the target. Search times and cumulative detection probabilities were measured for nine military targets in complex natural backgrounds.A total of 64 civilian observers, aged between 18 and 45 years, participate in the visual search experiment. The procedure of the search experiment is described in [6]. Search performance is usually expressed as the cumulative detection probability as function of time, and it can be approximated by [6] :

$$P_d(t) = \begin{cases} 0 & , t < t_0 \\ 1 - \exp\{-(t-t_0)/\rho\} & , t \geq t_0 \end{cases} \qquad (22)$$

where

- $P_d(t)$ is the fraction of correct detections at time t,
- t_0 is the minimum time required to response, and
- ρ is a time constant.

Fig. 5 shows the cumulative distribution functions corresponding to the search times measured for the target scenes used in the experiment here described. The overall difference between two of these functions can be measured by subtracting the area beneath their graphs. This operation corresponds to a Kolmogorov-Smirnov (K-S) test. To compare the relative distinctness of the targets in the different target scenes the curves are rank-ordered according to the area beneath their graphs. The resulting rank order for the target scenes is listed in the column with the header R_{P_d} in Table 1. These rank orders are adopted as the reference standard for the evaluation of the computational metric.
Targets that give rise to closely spaced cumulative detection curves which are similar in accordance with a K-S test, have similar visual distinctness. Fig. 5 shows that the target images in the experiment are clustered into a number of sets of targets with comparable visual distinctness: {16, 9, 37}, {6, 30, 26, 29}, and {3, 21, 11}. Consequently, rank order permutations of elements of the same cluster are not very significant, whereas rank order permutations of elements of different clusters are therefore significant.

6.3. Computational Target Distinctness

Let $C_n = \{\phi_{nj}\}$, with $n=1, 2, ..., N$, be the N natural clusters in *Active* produced by the RGFF model for the target image $t(x,y)$.

Figure 6: Target and empty images. Simple thresholding of the visual patterns produced by the target detetector on them.

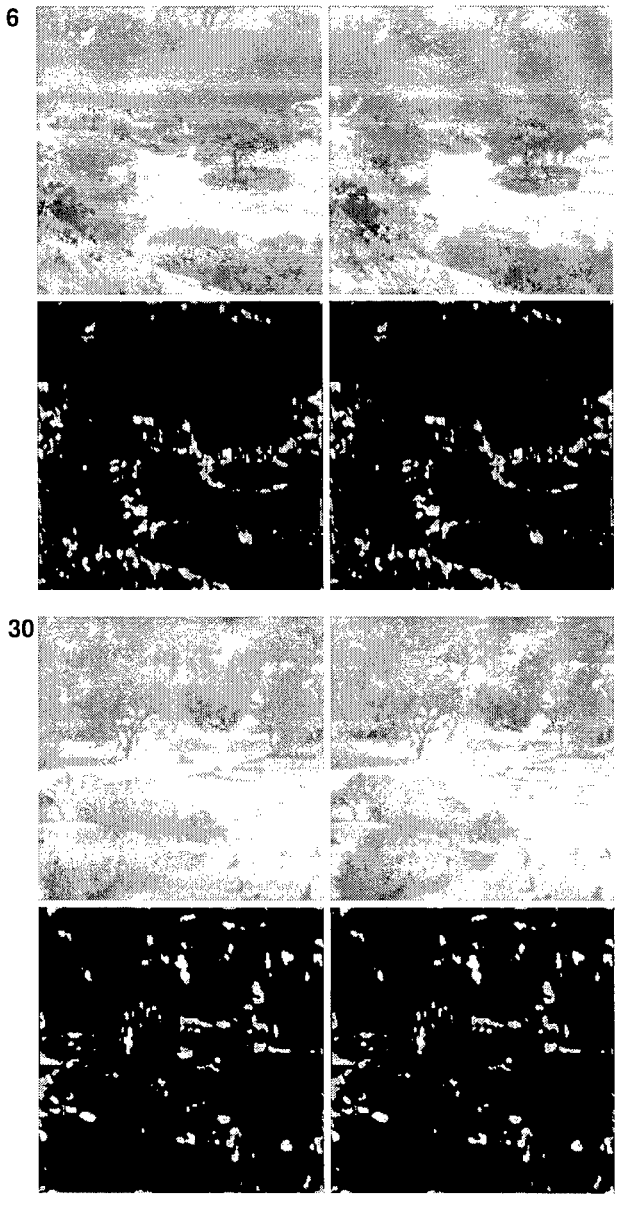


Figure 7: Target and empty scenes in the dataset, and the simple thresholding of the visual patterns produced by the target detector when it is applied on them. Simple thresholding was used to remove values which were present in the output of the target detector.

Let t_n represent the visual pattern segregated on the reference target image $t(x,y)$ by pooling the responses of filters in the natural cluster $C_n = \{\phi_{nj}\}$ as follows:

$$t_n = \left| \sum_j A_{n_j} \right| \tag{23}$$

where A_{nj}, denotes the original image $t(x,y)$ filtered through the logGabor ϕ_{nj} in C_n and passed through a non-linearity of the form:

$$tanh(z,\tau) = \frac{1-\exp\{-z\tau\}}{1+\exp\{-z\tau\}} \tag{24}$$

where τ is a gain term [16]. This nonlinearity enables the system to respond to local contrast over several log units of illumination changes.

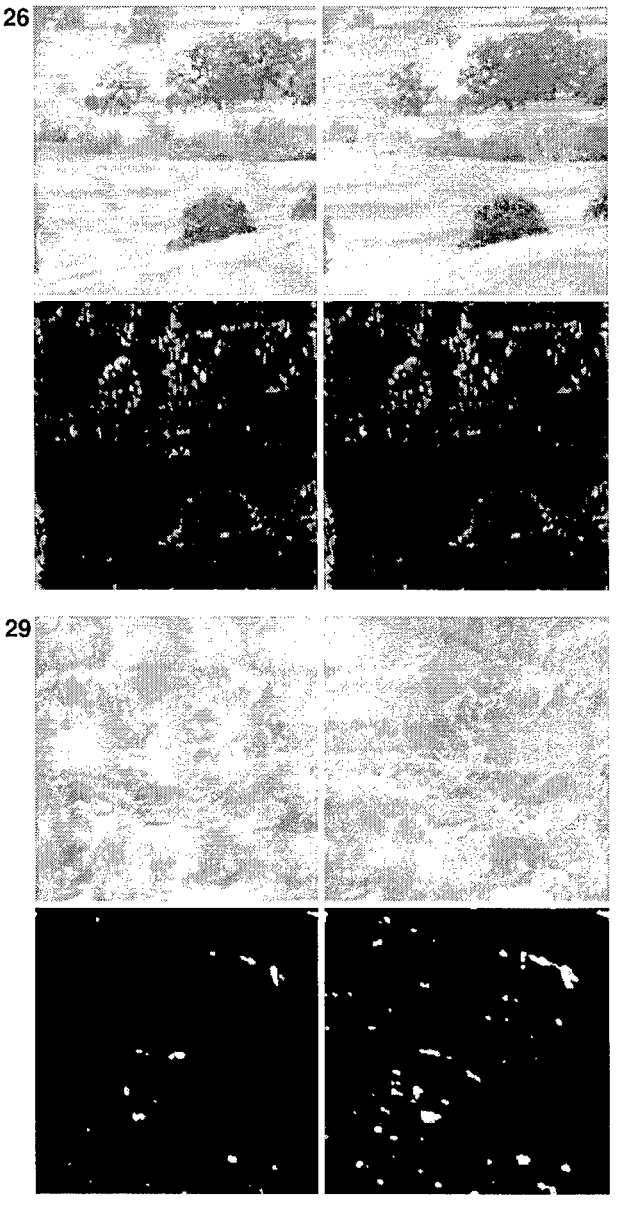


Figure 8: Target and empty scenes in the dataset, and the simple thresholding of the respective visual patterns by the target detector.

Therefore $t_1, t_2, \ldots, t_N$ represent a decomposition of the reference target image t into the set of its most significant visual patterns.

In order to compensate for the effect of image-to-image variations on the overall image light level, contrast normalization of each visual pattern is realized by dividing t_n by the sum of all filtered responses in *Active*, plus a saturation constant σ:

$$\frac{t_n}{\sigma^2 + \sum_i |A_i|} \tag{25}$$

where A_i denotes the original image $t(x,y)$ filtered through the logGabor ϕ_i in *Active* and passed through a non-linearity as given in equation (24).

Similarly passing the corresponding empty image $e(x,y)$ through the filters associated with each cluster C_n produced by the model on the reference image $t(x,y)$, results in a decomposition of e in $e_1, e_2, \ldots, e_N$.

Let $d_{VP}(t_n, e_n)$ be the difference between the visual patterns t_n and e_n, computed via the β-norm between their statistical structure over those pixels which form "fixation points" on t_n [11]:

$$d_{VP}(t_n,e_n)=\frac{1}{Card[FP(t_n)]}\left(\sum_{(x,y)\in FP(t_n)}\left|D\left[T^{t_n}(x,y),T^{e_n}(x,y)\right]\right|^{\beta}\right)^{\frac{1}{\beta}} \quad (26)$$

with $FP(t_n)$ being the set of fixation points for t_n; and $D[T^{tn}(x,y), T^{en}(x,y)]$ defining a normalized distance measure between the integral features $T^{tn}(x,y)$ and $T^{en}(x,y)$ computed on t_n and e_n, respectively. The default value of the exponent β in Equation (26) is 3.

Based on a definition of "visual pattern" as congruence in T across frequency bands, the differences between segregated visual patterns, $D_n=d_{VP}(t_n,e_n)$, $n=1,2,\ldots, N$, determine the overall distinctness between the reference target image t and the corresponding empty image e by using a simple decision rule:

$$VP_T(t,e)=\frac{1}{N}\sum_{n=1}^{N} D_n \quad (27)$$

A schematic overview of the VP_T distinctness measure is given in Fig. 10.

6.4. Relation between the computational and psychophysical target distinctness estimates

All the possible definitions of T were considered by recombining any subset of the next separable features:

- the phase T_1,
- the local energy T_2,
- the standard deviation of the local energy T_3,
- the local contrast of the local energy T_4, and
- the entropy of the local energy T_5.

For each specific definition of integral feature, noted as T, the notion of congruence in T across frequency bands was used to decompose the images into its visual patterns. The VP_T measure was then applied to quantify the visual distinctness of the targets. The subjective ranking induced by the psychophysical target distinctness was the reference rank order.

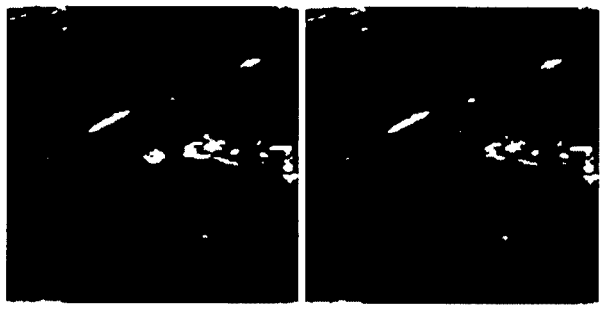
In order to study the efficacy of each definition T of integral feature for predicting target distinctness in a complex natural background, the fraction of correctly classified targets (with respect to the reference rank order) by the VP_T measure was computed on the dataset. Targets that give rise to closely spaced cumulative detection curves which are similar in accordance with a Kolmogorov-Smirnov test, have similar visual distinctness (Section 6.2.). Hence, the fraction of correct classification P_{CC} was defined as:

$$P_{CC}=\frac{Number\ of\ Correctly\ Classified\ Targets}{Number\ of\ Targets}$$

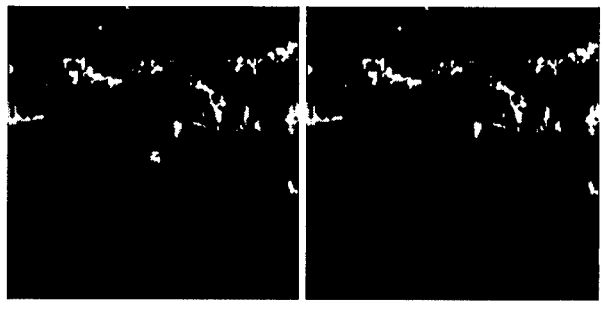
where rank order permutations of targets of the same cluster are insignificant (i.e., they are correctly classified by the metric), whereas rank order permutations of elements of different clusters are significant (the targets are then incorrectly classified).

The highest value of the fraction of correctly classified targets (P_{CC}=0.8) is obtained by the VP_T measure at $T=(T_1,T_2,T_3,T_4,T_5)$. Hence, the best definition of integral

3



21



11

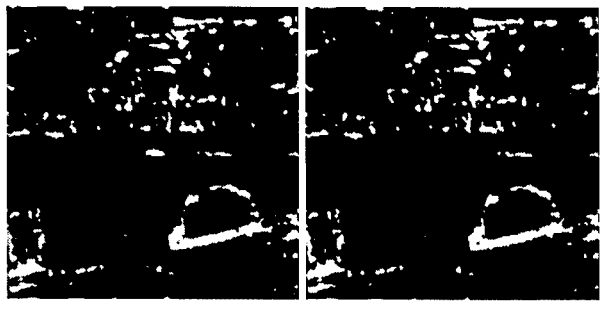


Figure 9: Target and empty images. Thresholding of the visual patterns produced by the target detector.

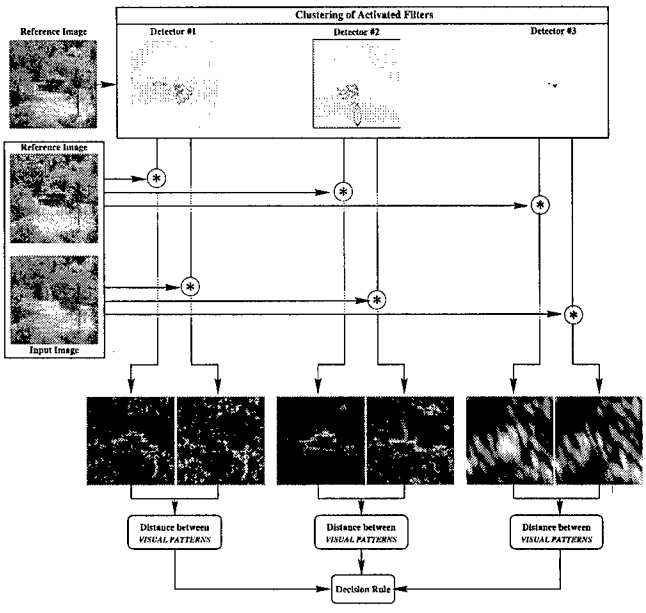


Figure 10: Schematic overview of the computational distinctness measure

feature for perceiving target distinctness on the dataset in this experiment, is $T = (T_1, T_2, T_3, T_4, T_5)$.
The comparative results of the *RMSE* metric and the VP_T measure based on the best definition of integral feature for predicting visual target distinctness are presented in Table 1. At the bottom of each of the columns is shown the respective fraction of correct classification. The reference rank order is listed in column 2.
The target distinctness values and the resulting rank order computed by the root mean square error (*RMSE*) metric are listed in column 3. The *RMSE* performs poorly, which is to be expected. Significant rank order permutations are displayed in boxes. The *RMSE* metric produces a rank order with five significant order reversals: targets 16, 26, 29, 3, and 11, are significantly out of order relative to the reference order induced by the psychophysical distinctness measure in column 2. The other targets have been attributed rank orders which do not differ significantly from the reference rank order. The *RMSE* yields a relatively low probability (P_{CC}=0.5). These results show that the *RMSE* metric appears not capable to rank order targets in the dataset with respect to their visual distinctness.
The target distinctness values and the resulting rank order computed by the $VP_{(T_1, T_2, T_3, T_4, T_5)}$ measure are listed in column 4. As noted above, this measure yields the highest probability (P_{CC}=0.8). This measure induces a rank order with two significant order reversals: targets 29 and 11 are ordered incorrectly. The other targets have been attributed rank orders which do not differ significantly from the reference rank order based on the psychophysical measure.
Summarizing, for the dataset in this experiment, the VP_T measure with $T=(T_1, T_2, T_3, T_4, T_5)$ appears to compute a visual target distinctness rank ordering that correlates with human observer performance.

7. CONCLUSION

Here a filtering technique was presented for the automatically learned partitioning of "visual patterns" in a digital image. Log-Gabor functions were adopted as an appropriate method to construct filters of arbitrary bandwidth. The novelty of our proposal lies in the definition of "visual patterns" as features which have the highest degree of alignment in statistical structure across different frequency bands. The interesting point is what kind of objects when imaged by cameras give rise to the visual patterns that the RGFF model segregates. They will be objects whose statistical structure across scales and orientations can be distinguished fairly well from the rest in a natural way. This limitation of the approach comes from the following assumption made in the clustering scheme of the Learning stage (Section[5]): the data set of activated filters has several separable clusters (e.g., elongated and non-piecewise linear separable groupings of arbitrary shape, dense and sparse natural clusters) and the membership is determined fairly well in a natural way by the data. The clarity of separation between clusters, as measured by a dissimilarity function, was the criterion by which they were derived. This assumption was needed to deal with several problems: (a) to overcome the lack of knowledge about the number and size of the clusters in the data, (b) to avoid the dependence of clustering on the initial cluster distribution, and (c) to find elongated and non-piecewise linear separable clusters, as well as to identify dense and sparse ones. In any case, the existence of natural clusters in the data is a very realistic assumption to many interesting applications. For example, because of the differences between the statistical structure across scales and orientations of targets and rural background in the application described in Section [6], the visual distinctness of a man-made object (a military vehicle) in a rural background can be determined in a natural way by the data.
Finally, a computational visual distinctness measure was presented that is computed from the image representational model based on visual patterns.It was applied to quantify the visual distinctness of targets in complex natural scenes. This measure that applies a simple decision rule to the distances between segregated visual patterns, was shown to correlate strongly with visual distinctness of targets in a dataset, as estimated by human observers.

Acknowledgments.

The authors thank Dr. Alexander Toet (TNO Human Factors Research Institute, The Netherlands) for providing us with image data, search times, and cumulative detection probabilities from search experiments made during the DISSTAF field test.
This research was sponsored by the Spanish Board for Science and Technology (CICYT) under grant TIC97-1150.

References

1. Field, D.J. "Scale-invariance and Self-similar `Wavelet' Transforms: an Analysis of Natural Scenes and Mammalian Visual Systems", in Wavelets, Fractals, and Fourier Transforms, Eds. M. Farge, J.C.K. Hunt, and J.C. Vassilicos, Clarendon Press, Oxford, pp. 151--193, 1993.
2. Morrone, M.C. and Burr, D. C. "Feature detection in human vision: A phase-dependent energy model." Proc. R. Soc. Lond. B, Vol. 235, pp. 221--245, 1988.
3. Field, D.J. "Relations between the statistics of natural images and the response properties of cortical cells", Journal of The Optical Society of America A, Vol. 4, No. 12, pp. 2379--2394, 1987.
4. Wertheimer, M. "Principles of perceptual organization," in Readings in perception, pp. 115--135, Van Nostrand, Princeton, NJ, 1958.
5. Lowe, D.G. "Three-dimensional object recognition from single two-dimensional images." Artificial Intelligence, Vol. 31, pp. 355--395, 1987.

6. Toet, A. "Computing visual target distinctness", TNO-report TM-97-A039, TNO Human Factors Research Institute, pp. 74, 1997.
7. Graham, Norma. Visual Pattern Analyzers. Oxford Psychology Series, No. 16, Oxford University Press, 1989.
8. Kovesi, P. "Image features from phase congruency", Technical Report 95/4, Department of Computer Science. The University of Western Australia. 1995.
9. Robbins, B. "The detection of 2D image features using local energy", D. Phil. thesis, Department of Computer Science, The University of Western Australia, 1996.
10. Fdez-Valdivia, J., Garcia, J.A., Martinez-Baena, J, and Fdez-Vidal, X.R. "The Selection of Natural Scales in 2D Images Using Adaptive Gabor Filtering", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, pp. 458--469, 1998.
11. Fdez-Vidal, X.R., Garcia, J.A., Fdez-Valdivia, J., and Rodriguez-Sanchez, Rosa. "The role of integral features for perceiving image discriminability", Pattern Recognition Letters, Vol. 18, pp. 733--740, 1997.
12. Treisman, A.M., and Gelade, G. "A feature-integration theory of attention." Cognitive Psychology, vol. 12, pp. 97--136, 1980.
13. Quick, R.F. "A vector-magnitude model of contrast detection". Kybernetik, 16, pp. 65--67, 1974.
14. Garcia, J.A., Fdez-Valdivia, J., Cortijo, F.J., Molina, R. "A dynamic approach for clustering data", Signal Processing, Vol. 44, pp. 181--196, 1995.
15. Toet, A., Bijl, P., Kooi, F.L., Valeton, J.M. "Image data set for testing search and detection models", TNO-report TM-97-A036, TNO Human Factors Research Institute, pp. 35, 1997.
16. Malik, J., and Perona, P. "Preattentive texture discrimination with early vision mechanisms", J. of Opt. Soc. Am. A, Vol. 7, No. 5, pp. 923--932, 1990.
17. Fdez-Valdivia, J., Garcia, J.A., and Garcia-Silvente, M. "An evaluation of the novel normalized-redundancy representation for planar curves", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 10, No. 7, pp. 769--789, 1996.
18. Witkin A. P. "Scale-Space Filtering". Proc. 8th Int. Joint. Conf. on Artificial Intelligence, Karlsruhe, West Germany, pp. 559--562, 1983.
19. Koenderink J.J., "The structure of images", Biological Cybernetics, Vol. 50, pp. 336--370, 1984.

APPENDIX

Let d_j'' be the second derivative of d_j computed as:

$$d_j''(i) = d_j(i) * \frac{d^2}{di^2} G_s(i)$$

with

$$G_s(i) = \frac{1}{\sqrt{2}s} \exp\left\{ -\frac{i^2}{2s^2} \right\}$$

and where d_j is convolved with the second derivative of the Gaussian at scale s, noted as $\frac{d^2}{di^2} G_s(i)$, to both smooth and differentiate the function.

The zero crossings of d_j'' correspond to positions at which the dissimilarity d_j undergoes a significant increment in its value. To locate the zero crossings marking a rise in d_j due to inter-cluster differences, the unwanted detail from intra-cluster differences must be removed by smoothing. The question is: how much smoothing should be performed? The derivative should be processed at the scale that best describes the increments in d_j due to inter-cluster differences, while removing spurious increments due to intra-cluster differences.Each interesting structure in d_j comes from a significant rise in d_j due to inter-cluster differences, and the best scale for describing the structrure should be based on its intrinsic redundancy across scales as follows [17].

Because structures of interest exist as significant entities over a certain range of scales [18,19], one expects to find some redundancy across the different scales if there exist significant structures in d_j. That is, a significant structure should have a greater similarity represented at its natural scales (the levels of resolution at which the structure can be perceived in d_j).

Two smoothed versions of d_j at successive scales will be correlated to the extent their structures are similar at the respective scales. And we can determine the degree of similarity by correlating the smoothed versions of d_j at successive scales.

Let $d_j^{s(l)}$ with $l=1,\ldots,L$ be the dissimilarity d_j smoothed by Gaussian kernels at several levels of smoothing $s(l)$ ranging in value from 1 to $s(L)$ and increasing by a constant of 0.5 from one level to the next. Then the normalized redundancy measure, denoted as $A(s(l))$, between two smoothed versions $d_j^{s(l)}$ and $d_j^{s(l+1)}$, at successive scales $s(l)$ and $s(l+1)$ can be computed by cross-correlating $d_j^{s(l)}$ and $d_j^{s(l+1)}$ as follows:

$$A(s(l)) = \frac{\left\langle d_j^{s(l)}, d_j^{s(l+1)} \right\rangle}{\| d_j^{s(l)} \| \; \| d_j^{s(l+1)} \|}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in the Hilbert space of measurable: square-integrable one-dimensional functions, and the norm (energy) of $d_j^{s(l)}$ is given by $\|d_j^{s(l)}\|^2$.

This function $A(s(l))$ returns a value measuring the relative redundancy between the respective smoothed versions at two consecutive scales. Given two smoothed versions, $d_j^{s(l)}$ and $d_j^{s(l+1)}$, at successive degrees of smoothing $s(l)$, $s(l+1)$ of a signal d_j, the value of the normalized function $A(s(l))$ at $s(l)$ is fairly small if any essential structure in $d_j^{s(l)}$has been removed from $d_j^{s(l+1)}$. Hence, each location $s(l)$ of local minima in $A(s(l))$ determines a significant scale for representing a structure of interest in d_j (i.e., a significant rise in d_j due to inter-cluster differences).

Consecuently, in order to locate the zero crossings of d_j'' marking a significant rise in d_j, the second derivative of d_j is then computed at the smallest scale from the set of locations $s(l)$ of local minima in $A(s(l))$. The derivative processed at the smallest significant scale, still describes the increments in d_j due to inter-cluster differences, while removing spurious increments due to intra-cluster differences.